Research Report

Precision and Volatility in School Accountability Systems

Walter D. Way

Research & Development

September 2006
RR-06-26

Precision and Volatility in School Accountability Systems

Walter D. Way

ETS, Princeton, NJ[1]

September 2006

As part of its educational and social mission and in fulfilling the organization's nonprofit charter and bylaws, ETS has and continues to learn from and also to lead research that furthers educational and measurement research to advance quality and equity in education and assessment for all users of the organization's products and services.

ETS Research Reports provide preliminary and limited dissemination of ETS research prior to publication. To obtain a PDF or a print copy of a report, please visit:

http://www.ets.org/research/contact.html

Abstract

This paper contrasts the positions of Kane and Staiger (2002) and Linn and Haug (2002) with that of Rogosa (2002, 2003a) on the volatility in test scores as measures of school growth. In particular, these researchers disagree on whether school growth can be measured reliably in school accountability systems. The different positions of these authors are examined in some detail. In addition, several issues related to their debate are examined in the context of the No Child Left Behind (NCLB) legislation, which has imposed a particular structure on the accountability systems currently in place in many states. Finally, some possibilities and speculations about the future of school accountability systems are provided in relation to the NCLB requirements and their impact.

Key words: School accountability, measurement of change, adequate yearly progress, vertical scaling

Introduction

Led by several states and bolstered by Title I amendments in 1994, the standards-based accountability movement greatly expanded during the 1990s. The movement had several characteristics, including greater focus on student performance, grading individual schools, the use of performance-level categories, public reporting, and rewards and sanctions assigned on the basis of performance (Fuhrman, 1999). Many states were not sufficiently equipped to resolve the many issues that arose when responding to the pressures to develop and implement accountability systems. Issues such as how to measure performance, defining satisfactory progress, the proper use of incentive systems, communicating results to the public, and the capacity of the states to remedy poor performance were not necessarily well understood by states as programs were set up. States varied in their implementation strategies and in the degree to which they were able to set up valid and credible programs.

By the early 2000s, several academic papers raised concerns about using state-based accountability systems to evaluate growth of student performance from year-to-year. In particular, Kane and Staiger (2002) and Linn and Haug (2002) argued that year-to-year changes in scores for groups of students are extremely unstable, and therefore, prone to misuse and misinterpretation. Using state-based data, Kane and Staiger (2002) illustrated the volatility of test score measures as tools for measuring change at the school level and suggested implications that were severe and alarming:

> Such volatility can wreak havoc in school accountability systems. To the extent that test scores bring reward and sanctions, school personnel are subjected to substantial risk of being punished or rewarded for results beyond their control. Moreover, to the extent that such rankings are used to identify best practice in education, virtually every educational philosophy is likely to be endorsed eventually, simply adding to the confusion over the merits of different strategies of school reform. (p. 236)

The cautions offered by these two papers and the analyses within them received a great deal of attention and endorsement from the measurement community. However, all researchers did not accept the conclusions from these papers. David Rogosa, a statistician and educational researcher with an extensive background in the measurement of growth, took strong exception to the positions taken by Kane and Staiger (hereafter referred to as K&S) and Linn and Haug

(hereafter referred to as L&H; Rogosa 2002, 2003a). Consider the following excerpt from the abstract of Rogosa (2003a):

> In this paper analytic results and artificial data examples are used to demonstrate that the methodology employed by both Linn-Haug and Kane-Staiger has absolutely no value and that their empirical demonstrations of volatility (or lack of stability) have no credibility. (p. 5)

A primary purpose of this paper is to evaluate the seemingly different positions on the volatility of school performance espoused by qualified and experienced researchers. In doing so, the arguments of L&H, K&S, and Rogosa will be examined in detail. In addition, several issues related to their debate will be examined in the context of the No Child Left Behind (NCLB) legislation, which has imposed a particular structure on the accountability systems currently in place in many states. Finally, some possibilities and speculations about the future of school accountability systems are provided, given the NCLB requirements and their impact.

The Case for Volatility in School Test Scores

Both K&S and L&H used several years of test data from different state assessments. K&S used test scores in reading and math for students in grades 3 through 5, obtained from the North Carolina Department of Instruction. They used the subset of matched records (about 65%of the records in any 1 year) in their analyses and matched students' test scores from 1 year to the next using information about date of birth, race, and gender These growth analyses were longitudinal in nature, in that the same cohorts of students were followed over multiple years. K&S also made use of school- and grade-level data on California's Academic Performance Index (API) scores in 1998 through 2000. API measures year-to-year growth based on successive cross-sectional cohorts of students within the same school. L&H used school-level fourth-grade reading data from the Colorado Student Assessment Program (CSAP) for successive cross-sectional student cohorts over a 4-year period.

Both K&S and L&H used reliability-based approaches to estimate the volatility of changes in test scores over time. K&S presented their approach in terms of variance components and characterized school test performance in terms of three factors: a permanent component that does not change before and after instruction, a persistent component that exists because of an innovation or an increase in instructional effectiveness (i.e., students improve because of some program or

practice), and a nonpersistent or transient component that is not repeated from year to year. In general, nonpersistent variance consists of sampling variation in students that occurs over time as different cohorts of students enter and exit a school. According to K&S, this variation can be substantial. Nonpersistent variance also includes miscellaneous group-level errors that can occur in the measurement process, such as bad weather or external disruptions on the day of the test, a severe flu season, and unique interactions between a group of students and a teacher in a particular year. K&S (2002) also presented data to support their argument that sampling variation is much greater for smaller schools than it is for larger schools. As a result, test scores will fluctuate more from year to year among small schools than among large schools. K&S caution that in states where rewards and sanctions are given to schools with the greatest and least growth, smaller schools are far more likely to be singled out simply because they are far more likely to experience extreme positive or negative growth due to sampling variation alone.

By representing a school's test performance as a sum of persistent and nonpersistent factors, K&S derived a formula to express the proportion of change in test scores that is attributable to nonpersistent factors. Assume a given school's test performance (S_t) consists of a permanent component (α), a persistent component (v_t), and a nonpersistent component, (ε_t). Then,

$$S_t = \alpha + v_t + \varepsilon_t,$$

where $v_t = (v_{t-1}) + u_t$, and u_t represents a new innovation or practice that contributes to student improvement.

K&S consider the correlation between change from year *t-1* to *t* (this year's change) and from year *t-2* to *t-1* (last year's change), and derive the following formula:

$$-2\rho = \frac{2\sigma_\varepsilon^2}{\sigma_u^2 + \sigma_\varepsilon^2},$$

where ρ is the correlation over schools between change this year and change last year, σ_u^2 is the variance in changes in test scores in 2 consecutive years due to persistent factors, σ_ε^2 is the variance in changes in test scores in 2 consecutive years due to nonpersistent factors, and $\sigma_u^2 + \sigma_\varepsilon^2$ is the total variance in the change in test scores from year to year According to this

equation, given an estimate of the correlation in changes in test scores over 2 consecutive years, the proportion of variance in changes due to nonpersistent factors can be estimated by multiplying that correlation by -2. Thus, if the correlation is close to -.5, nearly 100% of the changes that occur are transitory. If the correlation is close to zero, nearly 100% of the changes that occur are persistent. It should be noted that the equation only applies to correlations between 0 and -.5. K&S argued that this makes sense because there is no reason to assume that positive or negative changes in 1 year should lead to even greater positive or negative changes in the next. For the data sets that K&S analyzed, the correlations in changes over 1 consecutive year ranged from -.25 to -.43, implying that between 50% and 86% of the variance in these changes was transitory.

The approach taken by L&H to assess year-to-year changes in school-level scores differed slightly from the K&S approach. L&H computed change scores based on 2-year intervals by subtracting the percentage of students. classified in the Proficient or Advanced categories in 1997 from the corresponding percentage in 1999. Similarly, they created change scores by subtracting the 1998 percentage of students at the Proficient or Advanced levels from the 2000 percentage. L&H found that the correlation between changes from 1997 to 1999 and changes from 1998 to 2000 for 734 schools with CSAP scores in all 4 years was -.03 for the percentage of students in the Proficient or Advanced level. Using scores based on a weighted school-level index (based on assigning 1.5 times the percentage Advanced, plus 1 times the percentage Proficient, plus 0.5 times the percentage partially Proficient, minus 0.5 times the percentage unsatisfactory or with no test scores), a similar correlation of -.05 was found. L&H interpreted these correlations to illustrate the volatility in school change scores, in that knowing the 2-year change scores from 1997 to 1999 revealed nothing about the change from 1998 to 2000.

L&H offered similar cautions to those of K&S based on their analyses, suggesting that schools identified as outstanding or in need of assistance based on year-to-year changes in test scores are likely to be singled out based on chance factors:

> Because so much of the variability in school change scores is attributable to noise, it should not be surprising that schools identified as outstanding in one change cycle for achieving a large change in achievement are unlikely to repeat that performance in the next cycle. The converse is also true. Thus, schools that are identified as needing assistance in one cycle because they fell short of their change target, or even showed a

> decline, are unlikely to fall in that category the next change cycle. A consequence of this random fluctuation from one change cycle to the next is that the actions taken to assist schools in the latter situation may appear to be more effective than they actually are. Moreover, it is likely to be a mistake to assume that the practices of the schools recognized as outstanding are ones that should be adopted by other schools. (Linn & Haug, 2002, p. 7)

L&H also presented data on the changes in percentages of students in a school at the Proficient or Advanced level on the CSAP from 1997 to 1998 as a function of school size. They found that the variability of positive and negative change scores tended to decrease with school size, and concluded that a disproportionate number of small schools would be found with extreme changes (large or small) due to the greater variability in their change scores.

Both K&S and L&H offered suggestions that might improve the accuracy of results related to school growth. These included combining data across multiple grades, multiple subject areas, and/or multiple years. K&S proposed an application of empirical Bayes procedures, where estimates of school change would be a combination of the school's test scores; the state average; and the school's test scores from past years, other grades, or other subjects. Although this approach has the disadvantage of being less transparent to users and policymakers, both K&S and L&H suggested that the potential gains in precision would more than compensate for the increased complexity.

Rogosa's Position

The crux of Rogosa's objections to the K&S and L&H positions is their use of correlational analyses to demonstrate volatility in school test scores. This objection could be anticipated from Rogosa's earlier work. In a chapter discussing myths and methods for longitudinal research, Rogosa (1995) indicated that the proper basis for longitudinal research is collections of growth curves for individuals or groups. He emphasized that when there are small individual differences in change, low reliability of difference scores is an uninteresting consequence and unrelated to assessing the consistency of growth. Thus, Rogosa would view measuring growth as a matter of applying the right tools at the right level. From this perspective, it is not surprising that Rogosa (2003a) viewed K&S and L&H as applying inadequate tools for investigating consistency in improvement. He demonstrated this by setting up examples and data

simulations where the methods employed by K&S and L&H to assess the stability of school growth produced conflicting and irreconcilable results.

Rogosa's (2003a) first artificial example compared the improvement of five schools on the California API over 3 successive years, where a 50-point gain would be considered strong improvement. The year-to-year gain scores in this example are given in Table 1.

Table 1

Year-to-Year Gain Scores in Rogosa's First Example

	Year-to-year improvement		
School	Year 1 to 2	Year 2 to 3	Year 3 to 4
School A	40	50	50
School B	40	40	50
School C	50	40	40
School D	59	41	49
School E	45	45	45

Clearly, all five schools are showing notable improvement; yet, results of analyses following the methods of K&S and L&H will indicate extreme volatility for these data. In these data, the correlation between Year 1 to 3 and Year 2 to 3 is -.469. Using the K&S formula, the proportion of change in test scores due to nonpersistent variance is (-2)(-.469) or 93.8%. Using the L&H approach, we would calculate a Year-1-to-3 change (by adding the Year-1-to-2 change and the Year-2-to-3 change) and correlate that with a Year-2-to-4 change (the sum of Year 2 to 3 and Year 3 to 4). The result for the above example is a correlation of .0

In a second example, similar improvement is found across the five schools over 4 years. See Table 2. However, in this example the L&H methodology would again find great volatility in the school growth; whereas, the K&S methodology would conclude that all of the improvement is due to persistent change:

Table 2

Year-to-Year Gain Scores in Rogosa's Second Example

	Year-to-year improvement		
School	Year 1 to 2	Year 2 to 3	Year 3 to 4
School A	40	50	50
School B	40	40	50
School C	50	40	40
School D	50	50	40
School E	45	45	45

In this example, the correlation between Year 1 to 2 and Year 2 to 3 is .0. Thus, based on the K&S formula, 100% of the change in test scores is due to persistent variance. In contrast, the Year 1 to 3 and Year 2 to 4 changes in test scores are also correlated .0, so an analysis based on the L&H methods would conclude that there is great volatility in the data.

Rogosa (2003a) provided other examples where the volatility found in test scores differed according to the K&S and L&H methods. Because the K&S method only included 3 years of data in their analysis and the L&H method was based on 4 years, it was relatively easy for Rogosa to construct examples where the two approaches would disagree. But the point to take away from Rogosa (2003a) is his emphasis on assessing the year-to-year stability of school test scores and the year-to-year stability of differences between schools test scores. To Rogosa (2003a), the questions of interest in educational assessment and educational accountability have mostly to do with absolute (as opposed to relative) improvement:

- Did this school improve?
- Did this school improve *enough*?

Rogosa's position is consistent with the philosophy behind the No Child Left Behind Act, where the annual growth of schools is compared with specific growth targets. Rogosa (2003a) endorsed what he referred to as "qualitative aspects" of accountability systems, where growth targets are imposed on identified subgroups of students in addition to the entire school. He also provided examples of how schools can differ in their year-to-year improvement, as shown in Figure 1.

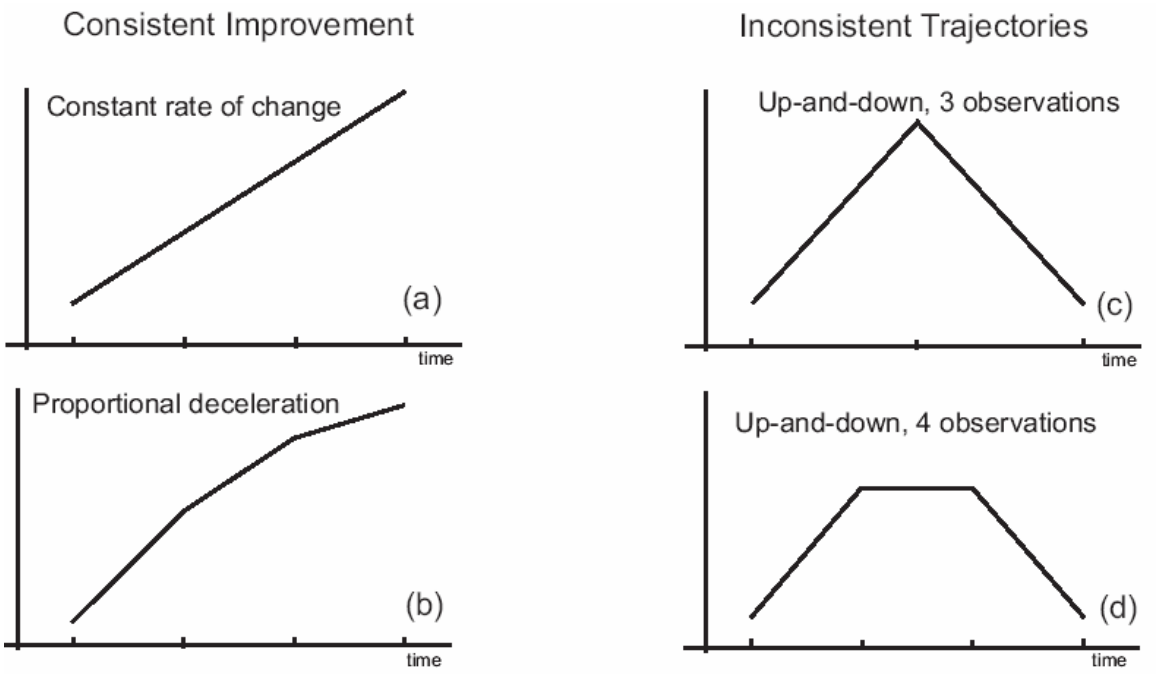

Figure 1: **Depictions of consistency of individual school improvement.**

Note. From *Confusions About Consistency in Improvement* (Figure 1.1) by D. Rogosa, June 2003, retrieved May 2, 2005, from http://www-stat.stanford.edu/%7Erag/api/consist.pdf. Copyright 2003 by David Rogosa. Adapted with permission.

When growth trajectories aggregate across individual schools, as in a statewide testing program, stability of improvement differs not only within a school across time but also between schools. Examples of the patterns that result are shown in Figure 2.

Rogosa (2003a) argued that the methods employed by K&S and L&H ignore the important features of school improvement data, which is the consistency of improvement for school trajectories. He emphasized the point by observing that although pattern (a) in Figure 2 represents an idealized situation of stable individual school improvement and stable between-school differences in improvement, the methods of both K&S and L&H would indicate 0% stability and 100% volatility in the observed growth, inviting the same kinds of warning messages that K&S and L&H provided based on their analyses.

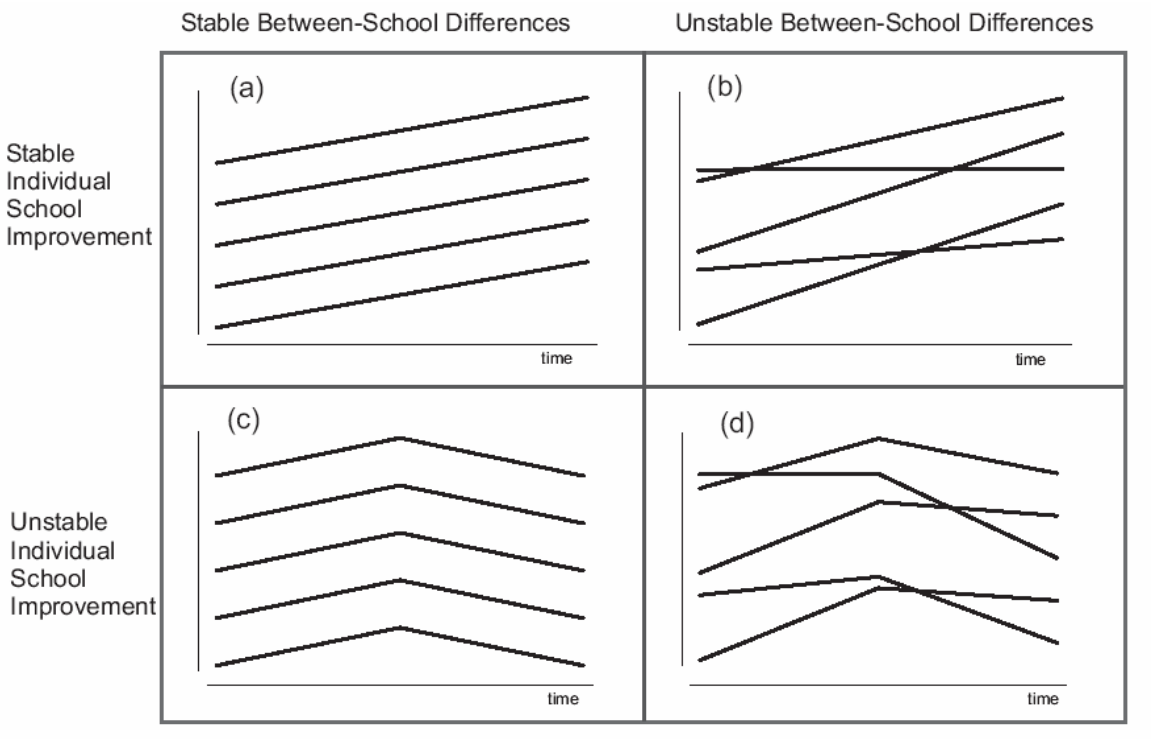

Figure 2. **Individual school improvement versus stability of between-school differences.**

Note. From *Confusions About Consistency in Improvement* (Figure 1.2) by D. Rogosa, June 2003, retrieved May 2, 2005, from http://www-stat.stanford.edu/%7Erag/api/consist.pdf. Copyright 2003 by David Rogosa. Adapted with permission.

Reconciling the Opposing Positions

Despite the apparent disjuncture between Rogosa's views on measuring school improvement and those of K&S and L&H, there are areas where their positions are not as much at odds as they appear. In attempting to reconcile Rogosa's position with those K&S and L&H, it is helpful to separate the general methodological approaches from the specific examples that are used to support the opposing positions. In presenting the argument that smaller schools are more likely than large schools to achieve awards for extreme year-to-year improvement, K&S (2002, Table 3) provided as an example summary information about gain scores in math in North Carolina between fourth and fifth grade by school-size decile. These data indicated that smaller schools were much more likely to achieve awards for improvement than larger schools. Rogosa (2002) refuted K&S by analyzing a second example referred to by K&S—that of year-to-year improvements in the California API—in great detail. It is important to note that there are huge

differences between the North Carolina and API examples. The API is based on a composite index of test results in four content areas that are aggregated across all grades in a school. Furthermore, the API data presented in the example required minimum school-level sample sizes of 100 students. Thus, the school-level API data were significantly more reliable than school-level Math scores in the K&S North Carolina example, and the API growth in the example was more accurately measured than growth in North Carolina scores example.

Rogosa (2002) admitted that smaller schools have an advantage over larger schools in showing API improvement due to greater variability in their scores, but argued that the greater variability also works the other way:

> A small school having made no real improvement has statistical variability as its friend, in that a false positive result may occur more often than for a large school. But a small school that has made substantial real improvement (which so far has been the more likely event) has statistical uncertainty as its foe, in that a false negative result may occur than for a large school. (p. 62)

Rogosa's (2002) analyses of the California API suggested only modestly higher false positive rates for smaller schools compared with larger schools. However, if Rogosa were to do similar analyses with the North Carolina math data, the K&S conclusions would likely be corroborated. That is, measuring school change for one content area and grade cohort will result in much greater volatility than measuring school change for a composite index based on several content areas combined across multiple grades. The impact of school sample sizes can be greater or lesser depending upon the design of the accountability system in place. There are also ways to minimize sample size as a factor in the determination of rewards and sanctions just as there are ways to control the volatility in year-to-year school-level test results within the accountability system.

In the end, it would seem that Rogosa, K&S, and L&H would all agree that small schools will have more volatile change scores than larger schools, but Rogosa would see this as something that can work both for and against small schools, as well as something that can be accounted for in a properly designed accountability system. A final point to consider with respect to smaller schools being identified for rewards and sanctions more often than large schools is that it would seem easier to engineer schoolwide improvement or suffer sharp schoolwide decline within a single year in smaller schools than in larger schools. This does not mean that

strong year-to-year growth for a small school is necessarily attributable to *random fluctuations* or that the small school is unworthy of recognition. But perhaps it does mean that the successful strategies leading to sizeable gains in the small school may not be as successful or even feasible in a larger school.

Rogosa's contrast of false positives and false negatives has additional implications when improvement relative to a goal is measured not only for a total group but, also, for various subgroups. Rogosa (2003b) commented on this issue as it relates to the Annual Measurable Objectives (AMOs) that have been set for NCLB accountability purposes in California for 2003–2005, that presenting calculations that reinforce the point that statistical variability in school and subgroup scores makes growth targets far more difficult than they seem, as each of the subgroups has larger uncertainty than the total scores for the school as a whole. This is consistent with similar points made by K&S (2002). Thus, both Rogosa and K&S would agree that requiring each subgroup to meet a specified growth target in addition to the total group will increase the likelihood of false negatives (i.e., the school's observed growth not reaching its accountability targets when in fact the school's true scores have). K&S view such requirements as presenting an undesirable disadvantage to racially integrated schools. However, Rogosa argues that such requirements are merely one of many other confounded factors (e.g., school size, demographics, school functioning, etc.) that can affect school accountability results, and he sees no basis in these requirements for claims of unfairness or disadvantage. To some extent, these differences in interpretation may be related to philosophical differences between K&S and Rogosa.

The Design of School Accountability Systems and NCLB

Most of the analyses and interpretations of K&S, L&H, and Rogosa preceded the passage of NCLB, which has had a dramatic impact on the design of state accountability systems. NCLB requires that by 2005–2006 that all states to assess mathematics and reading in grades 3–8 plus one high school grade. NCLB also requires states to set proficiency levels for each assessment and holds states accountable for making adequate yearly progress (AYP) each year through 2013–2014 (Roeber, 2003). The complexities of NCLB and the pressures of evolving existing accountability systems into NCLB compliance forced many states into making design choices that were expedient and likely to obtain approval. This seems to have resulted in a relatively limited range of implementations.

Under NCLB, each state must establish AYP targets. These are expressed as the percentages of students that reach Proficient or Advanced achievement levels in reading or language arts and mathematics, typically for a given year. NCLB specifies that each of at least 9 subgroups must all reach Proficient or above by 2013–2014, and states are responsible for setting baselines and interim annual targets. In addition, *safe harbor* provisions can apply at the school and district level. As spelled out in the final regulations, AYP is reviewed at two levels. The first level is a *status* review, and asks whether the school or district has met all of the AYP targets across measures and subgroups. The second level is *improvement*, and applies when one or more subgroup fails to meet the AYP status target. The essence of the safe harbor review is whether the percentage of students in that group who did not meet or exceed the Proficient level decreased by 10% from the preceding school year. For example, if 10% of a subgroup was Proficient or above in 2003 and 25% of that same subgroup was Proficient or above in 2004, the subgroup would fail a status review that called for percentage of Proficient or above students to be 35% or above for all subgroups. However, 90% of the subgroup was below Proficient in 2003 and only 75% of the subgroup was below Proficient in 2004. Since the percent Proficient or above for the subgroup increased 15% from 2003 to 2004, they would have met the safe harbor test (since 15% is greater than 9%, which is 10% of the 2003 percent below Proficient).

There are parts of the NCLB legislation where the possibility of more sophisticated methods for measuring school change can be inferred, but in general the provisions of the law make it extremely difficult to consider these methods. Marion et al. (2002) argue that a major aspect of defining a school accountability system requires asking what kinds of schools should be identified for improvement. The answer to this question can dictate the structure of the accountability system. For example, in *status models*, accountability is based on current achievement status (e.g., 50% of students are at or above Proficient in mathematics in 2004). In this model, the schools identified for improvement are those that do not attain the target status. *Successive groups models* evaluate accountability in terms of whether achievement is improving over time for the same grade (e.g., the percentage at or above Proficient in fourth-grade mathematics in 2004 compared with 2003). The California API, the analyses done by L&H using Colorado test data, and the safe harbor analyses that are part of most state AYP calculations are examples of the use of successive groups models. *A longitudinal model* evaluates progress from for the same group of students over time (e.g., average growth from third to fourth grade within a

school or district). *Quasi-longitudinal models* evaluate groups from year to year, but not all individuals are present in all calculations. In a longitudinal model, school effectiveness can be inferred by evaluating the achievement of individual students over time (e.g., individual growth curves).

These different models have different strengths and weaknesses. Status models are very easy to understand and school status scores are generally quite reliable. However, the models can only detect current level of achievement. Successive groups models are also easy to understand and compute, and they provide a mechanism for measuring school improvement. However, successive group models are susceptible to the sampling variation that occurs over time as different groups of students enter and exit a school. K&S (2002) and L&H (2002) both point out this limitation. Carlson (2002) summarizes the problem as follows:

> It is nearly unbelievable how little attention has been give to the validity of this approach, especially considering its widespread use. Many people assume that the mobility problem exists but that it isn't much of a problem for large schools or for schools with low mobility rates—or that it washes out with several years of data. There is some truth in all of these assumptions. What research is showing, however, is that it takes at least three or four years of data to draw a valid conclusion; secondly, that large schools are not immune to the effects of mobility or initial differences; and thirdly, that surprisingly low levels of mobility can render year-to-year, successive-group differences virtually uninterpretable. (p. 6)

Despite these clear technical limitations, Marion et al. (2002) note that the successive groups approach is the default model used by states as the basis for AYP definitions.

Longitudinal or quasi-longitudinal models include *value add*ed models, which have received a good deal of recent attention in the literature. McCaffrey, Lookwood, Koretz, Louis, and Hamilton (2004) provided a detailed review of the value-added-models literature. Although the technical details of these models are beyond the scope of this paper, several points related to their use for NCLB purposes are worthy of comment.

First, the use of longitudinal models, including value added models, typically involves vertically scaled achievement test scores. This practice has not been given a great deal of consideration in the literature, most likely because there is a longstanding history of norm-referenced achievement batteries that report and interpret vertically scaled scores. As a result,

casual observers of educational assessment assume that vertically scaled tests can be routinely incorporated into a state accountability system. However, AYP requirements under NCLB include an aligned system of academic content standards, academic student achievement standards, and assessments of student performance. In general, the development of standards-based tests is focused on within-grade content standards that are far less likely to have the kind of grade-to-grade continuity in content and difficulty specifications that norm-referenced test publishers have crafted into their instruments. Several recent studies have drawn attention to the issues related to using vertically scaled tests in value-added accountability systems (Martineau, 2006; Schmidt, Houang, & McKnight, 2005; Doran & Cohen, 2005). Wise (2004) has reported on an effort to evaluate the vertical alignment of state content standards.

Furthermore, the data collection designs that are needed for vertical scaling are not easily implemented in state testing programs, especially in those states where standards-based tests with separate within-grade scales are already in practice. These technical limitations have led some researchers to recommend against the use of vertical scaling for school accountability purposes. For example, Lissitz and Huynh (2003) proposed an alternative they referred to as "vertically moderated standards" as a mechanism that states might use to comply with NCLB legislation. In this approach, cut scores that define various proficiency levels are interpolated across grades to ensure that the achievement levels have the same (generic) meaning across grades, and that the proportions of students in each achievement level do not vary dramatically from grade to grade.

A second issue with the use of longitudinal or quasi-longitudinal models for NCLB accountability purposes is that No Child Left Behind also means *no child dropped from the longitudinal analysis*. That is, no NCLB-compliant assessment of school- or district-level growth can be made using only those students who can be matched from 1 year to the next. Thus, quasi-longitudinal approaches are as vulnerable to sampling error as the successive groups model. Although missing data can be identified and perhaps better accounted for in the analyses, interpretations of results will still be affected.

Finally, the infrastructure and data for tracking students longitudinally is not something that many states are currently capable of, and for those states it is unlikely that longitudinal accountability models will be feasible for some time. For example, although a school accountability task force in California recently argued strongly for a comprehensive data system

capable of tracking student information (Public Forum on School Accountability, 2003), it is unlikely that such a system will be in place in the short term.[2]

What Is Next for School Accountability Systems?

Given the structure imposed on school accountability systems by the NCLB legislation and the provisions for AYP, it seems likely that the NCLB systems implemented initially in various states will persist and be adopted by even more states. Even within the general NCLB structure, there are a myriad of design decisions that states need to consider. Marion et al. (2002) have provided a comprehensive analysis of these issues. One issue that has received recent attention is the reliability of school improvement results and their impact on NCLB accountability systems. Hill and DePascale (2002, 2003) have examined this issue in depth and many of their results and interpretations are consistent with those of K&S, L&H, and Rogosa. Hill and DePascale (2002) carried out extensive simulations involving random draws with replacement and Monte Carlo data generation to examine the reliability of accountability systems and the accuracy of decisions based on them. Their work focuses on classification errors and the impact of various design issues, such as sample size requirements for subgroup inclusion and the impact of school sampling issues (e.g., grade-to-grade variation is student ability generally or with respect to reading/English language arts versus mathematics). Their methods may also help to sort out questions about outcome variables, for example, whether the regulatory definition of safe harbor disadvantages states where growth in interim years may occur in ways that do not show up as increases in the percentages of students at or above Proficient. Many states defined performance levels prior to the enactment of NCLB, and the AYP requirements have introduced new and largely unknown implications for where the Proficient levels were set. In fact, some states have redefined Proficient in response to NCLB. Colorado now defines Proficient at the level that used to be labeled partially Proficient. The placement of the Proficient level within the state's test score distributions will affect whether and how AYP requirements are met, especially in the first few years of accountability.

It is likely that as more is learned about the full implications of state accountability system requirements under the NCLB legislation, states will begin to press federal officials to accept new options by which the AYP of schools and local educational agencies can be defined.[3] FairTest (2004) argues that standardized exams alone are inadequate measures if states seek to meet all the assessment requirements of the NCLB legislation. They point out that the language

in the law is ambiguous with respect to whether a state-administered system can have a mix of state-administered and local assessments and note that Maine and Nebraska are pursuing proposals with the U.S. Department of Education along these lines. Rabinowitz (2001) discusses the importance of balancing state and local assessments, describes attributes of models that blend state and local programs, and overviews steps in devising local systems in the context of state assessment programs. One possible model that might achieve this blend could have the state providing tools such as a *state item bank* and/or a *state toolkit of performance assessments* that could be offered over the Web and used by districts to build local assessments for formative and diagnostic purposes. This could provide all districts with common tools that could be used with minimal additional local development and could allow the state to set up some requirements for how they should be used for NCLB purposes. For districts with sufficient resources, individual plans involving district-developed assessments could be proposed and approved. The unifying component of the blended system would be the state content standards, which provide the common goals to align the local and state assessments. Roeber (2003) suggests that a model that mixed state and local district test models at alternate grades might be attractive. For example, the state would assess grades 4, 6, and 8 while the local district would assess grades 3, 5, and 7. However, he also warns that NCLB requires coherence in the skills that are measured as well as the information and decisions that result from the assessments. Thus, a blended system where a student could be judged Proficient on the local assessment in one grade but not on the state assessment at the next grade would not meet NCLB requirements. A more feasible alternative may be supplemental assessments and assessment tools, which may be sponsored by a state but that might be developed or purchased by local districts as they struggle to satisfy the state-imposed AYP requirements.

Another concern with the accountability system requirements under NCLB is that they tend to promote relatively narrow instruments and relatively limited information for teachers and administrators to use to improve instruction. Knowing the percentage of Proficient students in reading or mathematics in a school or that this percentage did or did not improve from last year provides little prescription to guide instruction or practice. Although standards-based tests may include subscores in different content areas, the subscores are typically not equated from year to year, and are either unreliable, too general to serve any real diagnostic purpose, or both. One way that states might improve the information that is used at the state or district level would be to mix

core standards-based assessments given to each student with matrix sampling. This would significantly broaden the sample of the state content standards assessed in a given year and make it easier to disclose items and field-test new items. Childs and Jaciw (2003) discuss pros and cons of partially-matrixed assessment designs and Dings, Childs, and Kingston (2002) report on simulations investigating the effects of partial matrix sampling on student score comparability in state assessments.

Whether state accountability systems will be able to expand their assessments beyond those that meet the letter of the NCLB requirements remains to be seen. Clearly, given the cost and complexity of customized state assessments, expanding them beyond NCLB compliance will be challenging to say the least. Yet in holding districts and schools accountable for student progress, state policymakers will be under pressure to look for ways to help them succeed. The good news for policymakers is that there are sophisticated assessment models and supporting technology to help in this effort. The challenge will be in harnessing these tools in ways that are compatible with the federal requirements, the purposes of the state assessment programs, and the needs of the local educators.

References

Carlson, D. (2002). *Focusing state accountability systems: Four methods of judging school quality and progress*. Retrieved May 15, 2002, from http://www.aceeonline.org/presentations/carlson_models.pdf

Childs, R. A., & Jaciw, A. (2003). Matrix sampling of items in large-scale assessments. *Practical Assessment, Research & Evaluation 8*(16). Retrieved February 19, 2004, from http://PAREonline.net/getvn.asp?v=8&n=16

Dings, J., Childs, R., & Kingston, N. (2002). *The effects of matrix sampling on student score comparability in constructed response and multiple-choice assessments*. Washington, DC: Council of Chief State School Officers.

Doran, H., & Cohen, J. (2005). The confounding effect of linking bias on gains estimated from value-added models. In R. W. Lissitz (Ed.), *Value added models in education: Theory and applications*. Maple Grove, MN: JAM Press.

FairTest. (2004). *Fact sheet: Flexibility needed. Possible state responses to new ESEA requirements*. Retrieved February 19, 2004, from http://www.fairtest.org/nattest/flexibiloty%20Needed.html

Fuhrman, S. (1999, January). *The new accountability* (The Consortium for Policy Research in Education Policy Brief RB-27). Retrieved May 15, 2004, from http://www.cpre.org/Publications/rb27.pdf

Hill, R. K., & DePascale, C. A. (2002, December*). Determining the reliability of school scores*. Washington, DC: Council of Chief State School Officers.

Hill, R. K., & DePascale, C.A. (2003, Fall). Reliability of No Child Left Behind accountability designs. *Educational Measurement: Issues and Practices, 22*(3), 12–20.

Kane, T. J., & Staiger, D. O. (2002). Volatility in school test scores: Implications for test-based accountability systems. In D. Ravith (Ed.), *Brookings papers on education policy, 2002* (pp. 235–260). Washington, DC: Brookings Institution.

Linn, R. L., & Haug, C. (2002). *Stability of school building accountability scores and gains* (CSE Tech. Rep. No. 561). Los Angeles, CA: Center for the Study of Evaluation.

Lissitz, R. W., & Huynh, H. (2003). Vertical equating for state assessments: Issues and solutions in determination of adequate yearly progress and school accountability. *Practical*

Assessment, Research & Evaluation, 8(10). Retrieved December 7, 2003, from http://PAREonline.net/getvn.asp?v=8&n=10

Marion, S., White, C., Carlson, D., Erpenbach, W., Hill, R., Rabinowitz, S., et al. (2002). *Making valid and reliable decisions in the determination of Adequate Yearly Progress*. Washington, DC: Council of Chief State School Officers.

Martineau, J. (2006). Distorting value added: The use of longitudinal, vertically scaled student achievement data for growth-based value-added accountability. *Journal of Educational and Behavioral Statistics, 31*(1), 35-62.

McCaffrey, D. F., Lookwood, J. R., Koretz, D., Louis, T. A., & Hamilton, L. (2004). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics, 29*(1), 67–101.

Public Forum for School Accountability. (2003, April). *A better student data system for California*. Oakland, CA: Public Forum on School Accountability.

Rabinowitz, S. (2001, December). Balancing state and local assessments. *The School Administrator Web Edition.* Retrieved February 19, 2004, from http://www.aasa.org/publications/saarticledetail.cfm?ItemNumber=2742&snItemNumber=950&tnItemNumber=951

Roeber, E. D. (2003, April). *Assessment models for No Child Left Behind* (Education Commission of the States Issue Brief: Accountability). Retrieved May 13, 2004, from http://www.ecs.org/clearinghouse/40/09/4009.doc

Rogosa, D. (1995). Myths and methods: "Myths about longitudinal research" plus supplemental questions. In J. M. Gottman (Ed.), *The analysis of change*. Mahwah, NJ: Erlbaum.

Rogosa, D. (2002, October). *Irrelevance of reliability coefficients to accountability systems: Statistical disconnect in Kane-Staiger "Volatility in school test scores."* Retrieved May 2, 2005, from http://www-stat.stanford.edu/~rag/api/kscresst.pdf

Rogosa, D. (2003a, June). *Confusions about consistency in improvement*. Retrieved May 2, 2005, from http://www-stat.stanford.edu/%7Erag/api/consist.pdf

Rogosa, D. (2003b, October). *California's AMOs are more formidable than they appear*. Retrieved May 2, 2005, from http://www-stat.stanford.edu/%7Erag/api/amo.pdf

Rogosa, D., & Haertel, E. (2003, July). *Deceived and confused: An attempt to reconcile the numbers in the Public Forum on School Accountability report, "A better student data*

system for California." Retrieved May 2, 2005, from http://www-stat.stanford.edu/~rag/api/Deceive.pdf

Schmidt, W., Houang, R. T., & McKnight, C. C. (2005). Value-added research: Right idea but wrong solution? In R. W. Lissitz (Ed.), *Value added models in education: Theory and applications*. Maple Grove, MN: JAM Press.

Wise, L. L. (2004, November). *Issues in assessment design, vertical alignment, and data management*. Paper presented at the CCSSO Meeting on the Use of Growth Models Based on Student-Level Data in School Accountability, Washington, DC.

Notes

[1] At the time this report was written, Walter Way was on staff at ETS. Currently, he is an employee of Pearson Educational Measurement.

[2] It is of some interest that the Public Forum on School Accountability document included a critique of the California API system, claiming that positive or negative API growth results were reversed in about 40% of schools in the Los Angeles Unified School District when growth was assessed in matched samples of students rather than assessed with a successive groups approach. Rogosa and Haertel (2003) refuted those claims, pointing out several inconsistencies and flaws in the data analyses. Based on Rogosa and Haertel's analysis, the disagreement rate in growth assessment was about 8%.

[3] As a case in point, the U.S. Department of Education recently announced that states may compete for inclusion in a pilot program that allows for incorporating growth-based accountability models as part of AYP. Eight states applied for the pilot program and two states, North Carolina and Tennessee, had their proposals accepted.